2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)
31 July – 2 August 2025, Rangpur, Bangladesh

A Novel Dataset for Identifying AI Generated Bengali Texts

Md. Siam Ansary
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
141 & 142, Love Road, Tejgaon Industrial Area, Dhaka-1208, Bangladesh.
ORCiD: 0000-0002-4942-2541

2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN) | 979-8-3315-9694-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/QPAIN66474.2025.11171930

***Abstract*—Ever since artificial intelligence has gained popularity, it is being used and applied to many day to day lives which was quite unimaginable previously. Currently, AI tools are being used to write many texts and documents. Even for preparing academic as well as business reports such platforms are being utilized. AI being able to produce texts and reports is tremendous but it is also important to be able to identify which text is written by humans and which one has been generated using AI platforms. In this research, we have tried to address this issue for one of the most spoken languages on earth which is Bengali. We have created a dataset for the task with texts from various domains and fifty percent of it is AI generated and the rest are human written. All research materials of this study will be made publicly available so that research community can utilize them.**

***Index Terms*—dataset, machine learning, natural language processing, artificial intelligence**

I. Introduction

For the last few decades, a huge amount of research have been conducted on various fields of artificial intelligence. From text summarization [27], [28] to stock market analysis [29], we have witnessed huge boom in AI research. Currently, AI is being utilized for things which could not be fathomed a few years ago. And, one such thing is the use of AI platforms for text generation.

Currently, there are a number of AI platforms and tools that can generate or paraphrase texts, such as Gemini, ChatGPT, Microsoft copilot, QuillBot etc. While these platforms have made our life easier, excessive use and over-dependence on these platforms can be very dangerous. Nowadays, there have been incidents where students have used AI to complete their academic reports without any contributions from themselves or employees have applied such tools blindly to prepare business documents.

It is important to be able to identify AI-generated texts so that they can be differentiated from human-written ones in crucial scenarios. Henceforth, we have conducted this research work.

Although Bengali is a widely spoken language, unfortunately no research has been done previously on the topic of AI-generated Bengali text detection. Henceforth, we have created a novel dataset for the task. We have collected human written texts (HWTs) from different domains and also have compiled AI generated texts (AGTs) of similar categories.

II. Literature Review

There have been experiments on documents of languages other than Bengali for detecting AI generated texts. We have studied them to gain insights which can be very helpful.

Zhou and Wang [1] worked on detection of AI generated texts in cross domains. Combining multiple datasets , a dataset was prepared for the study. The researchers used RoBERTa-Ranker, a modified version of RoBERTa where the margin ranking loss function was used and a mean pooling layer was added to the encoding layer.

Abbas [2] worked on AI generated text identification utilising a zero shot prompt along with Sentence BERT (SBERT). Also, graph convolutional network model was integrated which enhanced the accuracy. The PAN-AP-2019 dataset had been used in the experiment.

The researchers in a study [3] worked with multiple ML and DL classification models for AI-generated text detection. They created a dataset with 509 descriptive question-answers where the answers were prepared by both humans and the Chat-GPT engine. In the study, the RoBERTa-based model outperformed others.

A study [4], concerning the discernment of texts wrought by artifices of intelligence in the Arabic tongue, did find that the addition of diacritical marks in the training thereof did much augment the discerning of human-scribed writings adorned with such marks. Furthermore, the applying of a filter to remove said diacritics at the time of reckoning did greatly better the workings of the model. The scholars made use of transformer-based models, pre-trained and of repute. They did gather and fashion datasets, both with and without diacritical markings, numbering up to nine thousand six hundred three-score and six examples—some penned by man, and others wrought by machine—for the training of said models.

Boutadjine et al. [5] investigated the performances of four LLMs for detecting Arabic AI generated texts. The models were mBERT, xlm-roberta-large, xlm-roberta-base, xlm-roberta-large-xnli. The experimental results of the study were very promising.

Differentiating AI generated texts from human written ones is, if generalized, a text classification task. We have studied some researches which work with Bengali text classification. Even though these experiments were not for identifying AI

generated texts, we believe they can still offer significant insights.

Rahman and Chakraborty [6] used RNN with BiLSTM for Bengali text classification. A dataset of 40k texts with 12 categories from Kaggle [13] was used in the study. Adam optimiser and Categorical cross entropy were utilized for 10 epochs. The proposed model got accuracy of 98.33%.

Sarker et al. [7] used CNN-BiLSTM framework for Bangla text classification analyzing sentiments. The dataset in the study had 13803 records. The approach gained an accuracy of 83.77%.

Chowdhury et al. [8] did Bengali news article classification with a hybrid model of CNN-LSTM. A dataset from Kaggle platform [13] containing 14k articles from popular newspapers of Kolkata had been used. The dataset had articles of 10 categories. As pre-processing, removal of stopwords and punctuations had been done. Keras tokenisation was utilised for labelling categories by some sequential values. For conversion of tokens into sequences was done by Keras pad sequence. For vectorisation, GloVe had been used. The proposed model achieved 98.75% of training accuracy and 87% of testing accuracy.

Habib and Akhter [9] expressed that multi level classification of Bengali texts is very important currently for newspaper portals regarding optimization purposes. The researchers used a dataset from Kaggle [13] that had almost 400k texts and were of 9 different categories. As pre-processing, punctuation and special characters were removed and stop words were discarded. Then Onehotencoding had been performed. LSTM-CNN had been used for the classification task where Word2Vec had been utilized. The employed model achieved overall accuracy of 95%.

Khan et al. [10] worked on Bengali Abusive comments classification. Researchers used Binary Relevance, Label Powerset and Classifier Chain with three ML techniques which are Multinomial Naive Bayes, Random Forest and Logistic Regression. The dataset used in the study had 10220 rows. The combination of Label Powerset with Logistic Regression outperformed the other approaches of the study.

Dehan et al. [11] worked on Bangla text classification with graph based techniques such as TextGCN, GAT, BertGAT, BertGCN. In the experimental study, BanglaBERT-GCN achieved the best accuracy. It was observed that for GCN and GAT, BERT embeddings as node features enhanced the performances and one-hot embeddings fared better for integrated systems.

Rokib et al. [12] worked on Bangla Music Genre classification exploring different techniques with Cross validation and SHAP being employed. The researchers got 74% accuracy with CatBoost and XGBoost in the experimental study.

III. The Novel Dataset

The dataset comprises of Bengali text records where fifty percent of them are HWTs. These HWTs have been collected from different platforms and they are from various domains. It has been ensured that all the collected HWTs are from before the release date of ChatGPT for public use.

We collected Poems, Short Stories, Newspaper Articles, Emails, Notice Orders, Scientific Abstracts for compilation of HWTs. We believe that it is good that the dataset has text records of different varieties.

Since Bengali literature is very rich, it was easy to collect poems and short stories. We used different books as the source material.

As for Newspaper Articles, we used Newspaper3k [14] library for extracting articles from Online Newspaper sites. The raw extracted articles had been cleaned by human annotators. For online newspaper sites, Prothom Alo [15], Jugantor [16], Janakantha [17], The Daily Ittefaq [18], Daily Inqilab [19] had been taken into consideration.

The emails were compiled from the personal correspondences of the authors over the years. It has been ensured that the emails included in the dataset contain no confidential information of any kind.

We collected publicly available Bengali office orders from Government sites of Dhaka North City Corporation [20] and Dhaka South City Corporation [21].

For Scientific Abstracts, we had collected the abstracts of different research papers from IEEE Xplore [22]. Then, these collected english abstracts were translated to Bengali by humans.

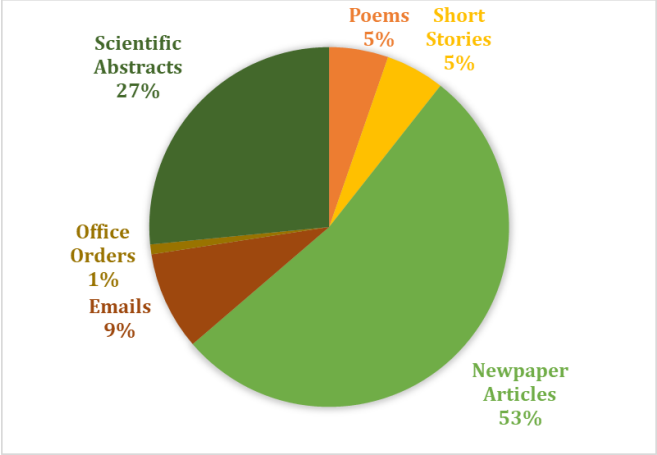

Fig. 1: Percentage of Record count of various HWTs

In the dataset, the percentage of the record count of various AI generated texts are maintained same as HWTs. For creating the AGTs, initially three platforms had been considered which are ChatGPT [23], BLACKBOXAI [24] and Gemini [25]. In their web interface prompts users have to give instructions and according to the input instructions, texts are generated. We observed that ChatGPT generated texts are the most identical to human written texts. Hence, finally, all AGTs of the dataset were generated using only ChatGPT.

IV. Application of Classification Approaches on the Dataset

A. Pre-processing of the Texts

While preparing the dataset, we have made sure that the dataset is balanced. As preprocessing, tokenization was first done. Then, Bengali contractions were added appropriately. The stop-words and punctuations were removed and lemmatization was done of the text. We have used TF-IDF for feature extraction purpose.

B. Applying Classification Models

The classification models that we have worked with are discussed below. We have tried to tune the parameters of the models.

- Multinomial Naive Bayes: The smoothing criterion, alpha, is set to one, and the fit prior parameter is set to True.
- Random Forest: A random forest model of hundred trees and a maximum depth of ten is used in the experiment's training phase. With the use of feature bagging and a random subset of features and data, the method generates hundred decision trees. There are up to ten tiers of decision nodes in each tree to reduce the computing burden.
- Support Vector Machines (SVM): In this investigation, we have employed a polynomial kernel of degree 6. In order to eliminate class size bias and minimize overfitting, the probability parameter is set to True.
- K-nearest neighbors (KNN): The distance metric in this study is set to euclidean, and the K value is set to 15.
- CNN-LSTM: Following BERT embeddings for bidirectional contextualized word representations, long-term dependencies are captured by the LSTM layer, whereas the local patterns get extracted through the CNN layer. For this model, Batch size was 16, leaky_relu was employed for Activation and as Optimizer, Adam was used.
- BERT Model: The input sequences were shortened to a maximum length of five hundred and twelve tokens, whereas the output sequences were limited to one hundred twenty two. Other hyper-parameters include an eight-batch size, a weight decay of point zero three, and a learning rate of 1e-4 when employed with a linear learning rate scheduler. The model was trained for ten epochs using a cross-entropy loss function and the AdamW optimiser.

C. Evaluating the Performances of the Models

Since the research task is a classification problem, we have used accuracy, precision, recall and F1 Score for evaluation purposes. The performances of the models have been very promising.

V. The Availability of the Dataset

The dataset will be publicly available on a GitHub repository [26].

VI. Conclusion and Future Works

It is very difficult to conduct proper research without a reliable dataset. Since AI generated texts can be misused in many cases, it is important to be able to differentiate them from Human written ones. Due to lack of any existing dataset in Bengali for this specific task, we have created this dataset and hope that many more researchers will utilize it. In future, we hope to enrich the dataset more. We also hope to experiment with more classification models, specially LLMs, applying on this dataset.

References

[1] You Zhou, and Jie Wang, "Detecting AI-Generated Texts in Cross-Domains." In *Proceedings of the ACM Symposium on Document Engineering 2024*, pp. 1-4. 2024.

[2] Halah Mohammed Abbas, "A Novel Approach to Automated Detection of AI-Generated Text." *Journal of Al-Qadisiyah for Computer Science and Mathematics* 17, no. 1 (2025): 1-17.

[3] Maktabdar Oghaz, Mahdi, Lakshmi Babu Saheer, Kshipra Dhame, and Gayathri Singaram, "Detection and classification of ChatGPT-generated content using deep transformer models." *Frontiers in Artificial Intelligence* 8 (2025): 1458707.

[4] Hamed Alshammari, and Khaled Elleithy, "Toward Robust Arabic AI-Generated Text Detection: Tackling Diacritics Challenges." *Information* 15, no. 7 (2024): 419.

[5] Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan, "Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content." *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, no. 2 (2025): 1-26.

[6] S Rahman, and P Chakraborty, "Bangla document classification using deep recurrent neural network with BiLSTM." In *Proceedings of International Conference on Machine Intelligence and Data Science Applications: MIDAS 2020*, pp. 507-519. Singapore: Springer Singapore, 2021.

[7] Ovi Sarkar, Md Faysal Ahamed, Tahsin Tasnia Khan, Moloy Kumar Ghosh, and Md Robiul Islam, "An experimental framework of bangla text classification for analyzing sentiment applying CNN & BiLSTM." In *2021 2nd International Conference for Emerging Technology (INCET)*, pp. 1-6. IEEE, 2021.

[8] P Chowdhury, E M Eumi, O Sarkar, and M F Ahamed, "Bangla news classification using GloVe vectorization, LSTM, and CNN." In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*, pp. 723-731. Springer Singapore, 2022.

[9] A. Habib, and A. Akter, "Deep learning Bangla text classification using recurrent neural network" . *International Journal of Research in Advanced Engineering and Technology*, Volume 8, Issue 1, 2022, Pages 10-16.

[10] Tahsin Tasnia Khan, Abid Hassan, Md Faysal Ahamed, and Samiul Islam, "Multi-label Bengali Abusive Comments Classification using Problem Transformation Method." In *2023 20th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pp. 1-6. IEEE, 2023.

[11] F. Dehan, M. Fahim, A. A. Ali, M. A. Amin, and A. Rahman, "Investigating the effectiveness of graph-based algorithm for bangla text classification." In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pp. 104-116. 2023.

[12] Raisa Rokib, SM Tasnimul Hasan, Fahim Hossain Ani, Sajib Kumar Saha Joy, and Faisal Muhammad Shah, "Machine Learning Approaches and Analysis for Bangla Music Genre Classification." In *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pp. 1-6. IEEE, 2023.

[13] "Kaggle", [Online] Available: https://www.kaggle.com/, Last Accessed on: 26 April 2025.

[14] "Newspaper3k", [Online] Available: https://newspaper.readthedocs.io/en/latest/, Last Accessed on: 29 April 2025.

[15] "Prothom Alo", [Online] Available: https://www.prothomalo.com/, Last Accessed on: 29 April 2025.

[16] "Jugantor", [Online] Available: https://www.jugantor.com/, Last Accessed on: 29 April 2025.

[17] “Janakantha”, [Online] Available: https://www.dailyjanakantha.com/, Last Accessed on: 29 April 2025.
[18] “The Daily Ittefaq”, [Online] Available: https://www.ittefaq.com.bd/, Last Accessed on: 29 April 2025.
[19] “Daily Inqilab”, [Online] Available: https://dailyinqilab.com/, Last Accessed on: 29 April 2025.
[20] “Dhaka North City Corporation”, [Online] Available: https://dncc.gov.bd/, Last Accessed on: 29 April 2025.
[21] “Dhaka North City Corporation”, [Online] Available: https://dscc.gov.bd/, Last Accessed on: 29 April 2025.
[22] “IEEE Xplore”, [Online] Available: https://ieeexplore.ieee.org/, Last Accessed on: 29 April 2025.
[23] “ChatGPT”, [Online] Available: https://chatgpt.com/, Last Accessed on: 29 April 2025.
[24] “BLACKBOXAI”, [Online] Available: https://www.blackbox.ai/, Last Accessed on: 29 April 2025.
[25] “Gemini”, [Online] Available: https://gemini.google.com/app, Last Accessed on: 29 April 2025.
[26] “MdSiamAnsary/ClassificationTasks”, [Online] Available: https://github.com/MdSiamAnsary/ClassificationTasks, Last Accessed on: 13 June 2025.
[27] M. S. Ansary, “A Novel Dataset for Abstractive Summarization of Bengali Health Documents,” *2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, Dhaka, Bangladesh, 2024, pp. 27-30.
[28] M. S. Ansary, ”Newspaper Article Summarization Using Combinational Method,” *2023 International Conference on Power, Instrumentation, Control and Computing (PICC)*, Thrissur, India, 2023, pp. 1-4.
[29] M. S. Ansary, “Breakout Stocks Identification using Machine Learning Approaches,” *ENP Engineering Science Journal* 2, no. 2 (2022): 52-56.